On The Prediction of Observables: A Selective Update

by

Seymour Geisser[1]
University of Minnesota

Technical Report No. 419

[1]This work was supported in part by NIH Grant GM25271

On The Prediction of Observables: A Selective Update

by

Seymour Geisser*

University of Minnesota

1. Introduction

During the last decade the predictive appraoch in statistical inference and decision has become more widely accepted as statisticians realize its pertinence and applicability to real problems. This is particularly true of Bayesians, because predictive distributions are a natural consequence of the Bayesian attitude, although, admittedly it has taken time for some Bayesians to adopt this view. Recently, Stigler (1982) claimed that the foundation of Bayes' original argument was to assume that if nothing were known about the observable event (number of successes in N trials) then each value could be presumed equally likely and equal to $(N+1)^{-1}$. In other words the focus was on the observables rather than on the unknown parameter (the postulated probability of success on any given trial). If Stigler's argument, based on his scrutiny of Bayes' Scholium, is valid then Bayes himself is the first Bayesian predictivist. One, however, may be puzzled as to why he couched his inference in terms of the parameter and not the chance of success of the (N+1)-st observation. Actually a calculation of the latter appears in the appendix which is due to Price, who communicated Bayes' posthumous essay. The calculation, as noted in the lectures of K. Pearson (1979), is somewhat obscure in the count of previous successes (whether it was one or two successes) and may not be consistent with the prior distribution of the parameter that was induced or assumed. Laplace (1774) correctly made the calculation and this came to be

*Supported in part by NIH Grant GM25271

known as his now notorious "law of succession". Is it then, that Bayes was not as much a predictivist as Stigler asserts? First, although it does not appear in the essay it does appear in the appendix, so it may have been suggested by Bayes. Secondly, the answer can better be determined by the specific problem that Bayes addressed. A ball was rolled on a unit square flat table and the horizontal coordinate of the final resting place was then assumed to be uniformly distributed in the unit interval. This perfectly reasonable assumption follows from the construction of the problem. A second ball is then rolled N times and one is informed as to the number of times the second ball came to rest to the left of the first ball without the actual horizontal coordinate of the first being disclosed. The problem is to infer the horizontal coordinate of the first ball. This is certainly predictivistic or observablistic inference in the broad sense in that it includes events that have occurred but whose values are unknown to the inferrer. A retrodictive inference is of the same nature as a predictive one for realizable situations. That Bayes himself did not specifically discuss the chance of a success on the toss of the (N+1)st ball for this problem reflects the fact that a different problem requires a different solution and hence should in no way mar Bayes' predictivistic credentials. However, when we go beyond this problem, and Stigler claims that Bayes did with regard to putting the prior on observables rather than on the parameter, then Laplace (1774) clearly undertook the next step at the posterior end, and even made the more general calculation of the chance that the next r out of M trials were successes, see also Condorcet (1786). Of course Laplace was responsible for greatly widening the scope and applicability of probabi-

lity calculations.

K. Pearson (1907, 1920) states that the fundamental problem of statistics was predictive and indicates how Laplace's type of calculation may be applied to a quantal response model.

Jeffreys (1939) extends this further by adopting a finite model, i.e. out of the total number of possible trials N+M he assumes that before any trials are made, the chance for the total number of successes is the same for every integral value from 0 to N+M. This yields the same result for the chance that the next r out of M trials are successes as the previous uniform prior assumption on the parameter itself.

A heated controversy broke out between Jeffreys (1932, 1933, 1934) and Fisher (1933, 1934) on a prediction problem. It concerned the probability that the third observation was included in the interval determined by the first two - all being independently and identically distributed. The controversy revolved more about the meaning of particular probability statements than about predicting. For a review of this controversy, see Lane (1981). Most of the problems Jeffreys deals with are physical measurement problems where the same true value is being measured imperfectly. Here the object is to infer limits on the true value so there is usually no need to predict future observations. However, he also derived the predictive distribution for (N+1)st observation given the first N are a random sample from a normal distribution.

Fisher (1956) devotes several pages of his book to prediction. He discusses the "Bayesian" calculation of obtaining r successes out of M future trials having previously observed s out of N and further demonstrates

the use of his fiducial argument for obtaining continuous predictive distributions. He makes some very penetrating remarks about the probabilistic prediction of observables and the capability of their verification as contrasted with probability statements about hypothetical parameters, as well as the connection between them. His views appear to be generally consonant with a predictivistic approach. But, because of a certain lack of clarity in his phraseology some of his assertions regarding predictions are capable of being interpreted in more than one way or perhaps misinterpreted.

Although, de Finetti and Savage, as far as I can discern, did not directly contribute to the methodology of statistical prediction, they have clearly provided the major philosophical underpinning for the observabilistic or predictivistic view. In fact, next to Bayes' theorem, de Finetti's exchangeability theorem is second to none in its importance for the subjectivistic view. It is also useful in demonstrating that a good deal of parametric inference can be viewed as a special or limiting case of observabilistic inference. The distribution of unobserved but realizable values or particularly important functions of them, as modified by values observed, is clearly what the purview of the bulk of inference should be.

It is interesting to observe that, philosophy notwithstanding, most of the methodology produced by Bayesians whether of the parametric or predictive variety tends to be more relaxed and approximate in its construction than one would anticipate given a strict subjectivistic viewpoint. The methods tend to be more pragmatic than subjective and tend to conform to the spirit of the Bayes/Non-Bayes compromises of Good (1965). It also concords with the view (shocking to some) that a parameter is often an

unobservable construct of an approximate model mainly devised to facilitate the prediction of future observations. Hence a prior distribution for the parameter can be, to a degree, a matter of convenience.

In the next section we sketch out some of those areas for which predictive methods have been devised during the last twenty years. Featured in the third section are some recent predictive developments in regard to influential observations, data consistency and model checking. The last section discusses some results in predicting the number of future observations out of a total that lie in a chosen set. The latter can be considered a continuation, of sorts, of the Bayes-Laplace-Pearson-Jeffries-de Finetti calculations.

2. Applications of Bayesian Predictivism

Forecasting in time series is a very natural enterprise and consequently a well developed field as attested to by the voluminous literature. We mention here only a few of the more notable statistical books that stress time series prediction; Wold (1938), Wiener (1949), Yaglom (1962), Whittle (1963), and Box and Jenkins (1970). Most of the orientation is non-Bayesian, and a good deal of effort is devoted to estimation e.g. Anderson (1971). Wold (1959) veered away from parametric estimation to a predictivistic approach for econometric data - but this apparently had no appreciable effect on other non-Bayesian workers. From here on I will stress only areas other than time series, where statistical prediction was evidently not as natural to the developers of statistical methodology and the critical focus was on the estimation of parameters and the testing of hypotheses about them.

The use of predictive distributions in Bayesian classification and discrimination problems was noted and elaborated upon by Geisser (1964, 1966, 1968) and in multivariate normal linear regression, Geisser (1965). Aitchison and Sculthorpe (1964) discuss prediction from a decision perspective mainly with regard to tolerance and coverage problems, see also Guttman and Tiao (1964). Dunsmore (1968, 1969, 1974) used the predictive approach to problems of calibration, life testing, regulation and optimization and Guttman (1967) to goodness-of-fit problems. Roberts (1965) and Geisser (1971) point out various non-traditional predictive areas, where Bayesians were using predictive distributions, such as classification, discrimination, certain hypotheses testing problems, design, sample size determination, sample surveys, goodness-of-fit, without necessarily acknow-

ledging the fact. Geisser (1971) also points out that problems of ranking, selection and comparison handled, even by Bayesians, from an estimative viewpoint are better executed predictively. In short, predictive distributions had been grossly underutilized by Bayesians!

Prediction in growth curve situations were developed by Geisser (1970), Lee and Geisser (1972, 1975), and Fearn (1975). Finally, Aitchison and Dunsmore (1975) published the first text in statistical prediction covering a substantial number of predictivistic topics. For a detailed review, see Geisser (1976). The predictive approach has also been utilized by Akaike (1978) and Geisser (1979) in an attempt to induce objective prior distributions for parameters.

On the non-Bayesian side, except for the previously alluded to comments by Fisher and the early frequentist work by Wilks (1942) on distribution-robust tolerance regions, little attention had been paid to the predictive approach. A summary of classical frequentist tolerance procedures is given by Guttman (1970).

A new predictive impetus in low structure paradigms (and to problems with varying structure) derives from the work of Geisser (1974, 1975a, 1975b, 1979, 1980a, 1980b, 1980c, 1981), Stone (1974a, 1974b, 1977), Butler and Rothman (1980), Wahba (1977) and deWaal et al (1981) who use sample reuse procedures to make predictions, select models, estimate densities, and modify classification techniques, among other things.

That in the softer social, biological and engineering sciences prediction should always have been a crucial factor is well known, but recently Jaynes (1980) has made a most convincing case for a predictivistic version

of statistical mechanics that is directed towards what he regards as the critical question: "Given the partial information that we do, in fact, have, what are the best predictions we can make of observable phenomena?"

In the next two sections some recent developments in the use of predictive distributions are presented and elaborated upon in some detail.

3. Influential Observations, Data Consistency And Model Checking

3.1 Influential Observations

How observations effect the estimation of certain parameters has been the focus of much attention for the last ten years. Cook (1977, 1979), Cook and Weisberg (1980, 1982), Andrews and Pregibon (1978), Hoaglin and Welsch (1978), Belsley, Kuh and Welsch (1980) have considered the problem of detecting observations which are influential in the estimation of regression parameters. Johnson and Geisser (1981, 1982, 1983) have considered the problem of influential observations using a Bayesian approach and deriving methods both for the estimation of parameters and the prediction of future observations. Methodology is developed to ascertain how an observation (or set of them) influences the posterior distribution of a set of parameters of interest or the predictive distribution of a future set of observables. The approach is to compare the posterior (predictive) distribution of the parameters (future observables) with and without the set of observations whose influence is to be determined. Indicators of the discrepancy between the two distribution functions such as the Jeffreys-Good-Turing-Kullback-Leibler information measures, are used, c.f. Kullback and Leibler (1951). In particular, for estimation purposes, one computes the posterior marginal distribution of the set of parameters of interest and applies the previous notions to them.

Although the general Bayesian methodology set out by Johnson and Geisser (1981, 1982, 1983) can be used in any situation requiring it, the most important applications have been in regression analysis. In regression problems the work of Cook (1977, 1979) is especially note-

worthy. He proposed a statistic as an indicator of the influence that an observation has with regard to the estimation of a set of regression parameters. And as the regression problem is still of the widest interest, but by no means the only important application, we shall give prominence to it in the discussion here.

Consider a normal linear regression situation where

$$\begin{aligned} Y &= X\beta + e, & e &\sim N(0,\sigma^2 I) \\ Y' &= (Y_1,\ldots,Y_n), & e' &= (e_1,\ldots,e_n) \\ x_i' &= (1,x_{i1},\ldots,x_{ip}), & \beta' &= (\beta_1,\ldots,\beta_p) \end{aligned} \tag{3.1}$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} \tag{3.2}$$

with assumed prior density for β and σ^2, say

$$g(\beta,\sigma^2) \tag{3.3}$$

The first step in assessing the influence of individual observations with regard to the estimation of β alone, say, is the computation of

$$p(\beta) = p(\beta|y,X) \propto \int L(\beta,\sigma^2|y,X)\ g(\beta,\sigma^2)\ d\sigma^2 \tag{3.4}$$

and

$$p_{(i)}(\beta) = p(\beta|y_{(i)},X_{(i)}) \propto \int L(\beta,\sigma^2|y_{(i)},X_{(i)})\ g(\beta,\sigma^2)\ d\sigma^2 \tag{3.5}$$

where y the observed value of Y is decomposed such that $y_{(i)}$ is y with y_i deleted and similarly $X_{(i)}$ is X with the i^{th} row deleted and $L(\cdot)$ is the likelihood function. Next we compute one of the information measures, say,

$$I_i(\beta) = I(p_{(i)}, p) = E[\ln p_{(i)}(\beta) - \ln p(\beta)] \tag{3.6}$$

where, by definition, the expectation is taken with respect to the first density. All of the observations y_i are then ordered according to $I_i(\beta)$, the larger this value the more influential is y_i. Of course one can include σ^2 as well and if this is the goal we can calculate the joint information measure

$$I_i(\beta, \sigma^2) = I_i(\sigma^2) + E[I_i(\beta|\sigma^2)] \tag{3.7}$$

where

$$I_i(\sigma^2) = E[\ln p_{(i)}(\sigma^2) - \ln p(\sigma^2)], \tag{3.8}$$

and $p_{(i)}(\sigma^2)$ and $p(\sigma^2)$ refer to $p_{(i)}(\sigma^2|y_{(i)}, X_{(i)})$ and $p(\sigma^2|y,X)$ respectively; similarly

$$I_i(\beta|\sigma^2) = E[\ln p_{(i)}(\beta|\sigma^2) - \ln p(\beta|\sigma^2)] \tag{3.9}$$

and $I_i(\beta|\sigma^2)$ is averaged over the density $p_{(i)}(\sigma^2)$. This partition often helps to pinpoint the sources of influence. Details of this approach with examples have been worked out for the multivariate general linear model by Johnson and Geisser (1981).

When the stress is on prediction as it often should be in regression problems, it is necessary to calculate the predictive distribution of Z,

the $m \times 1$ future vector to be observed for a given W, an $m \times p$ matrix, i.e.

$$Z = W\beta + e^* \qquad e^* \sim N(0, \sigma^2 I) \tag{3.10}$$

with and without y_i. Consequently,

$$f_{(i)}(z) = f_{(i)}(z|W, y_{(i)}, X_{(i)}) = \int f(z|W,\beta,\sigma^2)\ p_{(i)}(\beta,\sigma^2)\ d\beta d\sigma^2 \tag{3.11}$$

$$f(z) = f(z|W, y, X) = \int f(z|W,\beta,\sigma^2)\ p(\beta,\sigma^2)\ d\beta d\sigma^2 \tag{3.12}$$

One then calculates

$$I_i(Z) = E[\ln f_{(i)}(Z) - \ln f(Z)] \tag{3.13}$$

as the predictive influence function (PIF) which is used for the relative assessment of influence of the y_i's with regard to predicting a future set of values at W. This is useful only in as far as prediction at W is at issue. If W is unknown but can be assigned probabilities then this can be incorporated into the assessment. If this is not the case, it has been found useful to set W = X, i.e. to essentially ascertain the effect of predicting back on the original set of independent variables as indicative of an overall assessment. The details of this procedure are given by Johnson and Geisser (1982, 1983).

To demonstrate the calculations in the simplest fashion, we use the vague prior for β and σ^2,

$$g(\beta,\sigma^2) \propto \frac{1}{\sigma^2} . \tag{3.14}$$

Let x_i' be the i^{th} row of X then define

$$v_i = x_i'(X'X)^{-1}x_i, \qquad (N-p)s^2 = (y-\hat{y})'(y-\hat{y}),$$

$$\hat{\beta} = (X'X)^{-1}X'y, \qquad \hat{y} = X\hat{\beta}, \qquad \hat{y}_i = x_i'\hat{\beta}, \tag{3.15}$$

$$t_i^2 = \frac{(\hat{y}_i - y_i)^2}{(N-p)s^2(1-v_i)}, \qquad D_i = \frac{v_i(N-p)}{(1-v_i)p}\, t_i^2.$$

(The statistics D_i was defined by Cook (1977) as a measure of the influence of y_i on the estimation of the set of regression coefficients β.)

Using these results we can calculate the various measures of influence previously defined. First we obtain $2I_i(\beta,\sigma^2)$ which is the sum of the two expressions,

$$2I_i(\sigma^2) = C + (N-1-p)t_i^2(1-t_i^2)^{-1} + (N-p)\ \ln(1-t_i^2) \tag{3.16}$$

$$2E[I_i(\beta|\sigma^2)] = K + (N-1-p)\ \frac{v_i t_i^2}{(1-v_i)(1-t_i^2)} + \frac{v_i}{1-v_i} + \ln(1-v_i) \tag{3.17}$$

where C and K are constants independent of the deleted observation. Although an explicit expression for $I_i(\beta)$ is not obtainable the following approximation, based on a "best" scaled multivariate normal approximation to a multivariate student distribution, should be more than adequate

$$2\hat{I}_i(\beta) = \frac{(N-p-2)v_i}{1-v_i}\ t_i^2 + \left[\frac{v_i}{1-v_i} + \ln(1-v_i)\right] + p\left[\frac{N-p-2}{N-p-3} + \ln\frac{N-p-3}{N-p-2} - 1 - \ln(1-t_i^2) - \frac{t_i^2(N-p-2)}{N-p-3}\right]$$
$$- \frac{v_i}{1-v_i}\left[\frac{(N-p-2)(t_i^2-1)}{(N-p-3)} + 1\right]. \tag{3.18}$$

It is to be noted that the lead term of the above expression, which reflects the effect of fit times leverage, is proportional to the influence function D_i proposed by Cook (1977). The other components

reflect the effect on the variation and volume of the posterior distribution of β.

For the predictive influence function a similar "best" multivariate normal approximation to a multivariate student distribution is utilized. This results in

$$2\hat{I}_i(Z) = \frac{(N-p-2)v_i t_i^2 (N-p-4)}{2(1-v_i)(N-p-3)} + \left[\frac{v_i(N-p-2)}{2(1-v_i)(N-p-3)} - \ln\left(1+\frac{v_i}{2(1-v_i)}\right)\right] + N\left[\frac{N-p-2}{N-p-3}(1-t_i^2) - \ln\frac{N-p-2}{N-p-3}(1-t_i^2) - 1\right]. \tag{3.19}$$

Again, the first term is proportional to Cook's D_i and measures mainly the change in location of the center of the predictive distribution according to some metric (lack of fit) multiplied by a measure of leverage. The second component is essentially a measure of leverage while the third component reflects the effect on the observational error. From the predictive viewpoint, an influential observation will be distantly observed, exhibit lack of fit and significantly alter the volume of the predicting ellipsoid.

For a very large sample it is clear that Cook's statistic, D_i, is adequate for the influence measure with respect to β, since the first term in $\hat{I}_i(\beta)$ is of an order of magnitude larger than the subsequent terms. However, as the sample grows, the joint influence measure $I_i(\beta,\sigma^2)$ depends not only on D_i but a convex function of $(N-p)t_i^2$ as well. With regard to $\hat{I}_i(Z)$, the predictive influence function, the situation is the same with greater relative emphasis on the convex function of $(N-p)t_i^2$.

Moreover, for small and moderate sample sizes all of the influence

measures can differ considerably from D_i in the relative influence assigned to the observation.

For the details of applying these notions to univariate and multivariate general linear models including the influence of subsets of size k, alternative information measures and the use of conjugate prior distributions see, Johnson and Geisser (1981, 1982, 1983). Particular data sets are analyzed in these papers to demonstrate the methods as well as to pinpoint differences in influences as compared with the method of Cook.

Other paradigms are currently begin examined such as growth curves, time series, classification and discrimination and censored data situations to ascertain the influence that particular observables have on the resulting inferences and decisions that are to be made.

3.2 Data Consistency

Once relatively influential observations are detected, the data analyst is concerned with the source of the observation's influence. Are there substantive grounds for deletion of the observation? Upon investigation it might be ascertained that the value was inaccurately recorded or some departure from the experimental protocol or other adverse condition occurred that would render the observation defective. When there is no reason to suspect an observation's defective nature, other possibilities in regression situations are that it was distantly observed or inconsistent with the rest of the data in terms of the adequacy of the model. One can determine whether it was distantly observed simply enough by the relative values of v_{ii}. To check for its consistency with the rest of the data, an influential observation can be subjected to a predictive significance test. Ait-

chison and Dunsmore (1975) refer to it as the atypicality index of a new observation and discuss its use in several applications. On the assumption that the observations $y_{(i)}$ are mutually consistent in regard to the model, a small value of

$$\Pr\{Z:\ f_{(i)}(Z|X_i,X_{(i)},y_{(i)}) \leq f_{(i)}(y_i|X_i,X_{(i)},y_{(i)})\} \tag{3.20}$$

could cast doubt on whether y_i is consistent with $y_{(i)}$. Here the future value Z is random and $y_{(i)}$ is fixed as opposed to the sampling situation where both are assumed random. This calculation, of course, presumes that y_i was chosen before the value was actually observed. If in fact y_i was so chosen because it had maximum influence, then, presumably, one ought to condition on this fact, but this greatly complicates the exact calculation of the "significance" value.

It is also possible to use, as a diagnostic,

$$d_i = f_{(i)}(y_i|X_i,X_{(i)},y_{(i)}) \tag{3.21}$$

Geisser (1980) to search for possible observations that are inconsistent with the rest. Although d_i is obviously affected by transformations on y_i and $y_{(i)}$, this will not be a serious matter, see the discussions by Box (1980) and Stigler (1980). When an observation appears to be inconsistent with the rest, the data analyst may have to decide whether to retain or delete the observation or perform analysis with and without the offending observation. Certainly if the observation's influence is minimal it is of little consequence as to which alternative is chosen.

Thus far only one model has been entertained and a search for single observations or small subsets that were inconsistent with that model were conducted. For the linear regression problem of section 3; using the vague prior (3.14), it is easy to show that (3.20) is equivalent to

$$\Pr\left\{\frac{t_i^2}{1-t_i^2} \leqslant F(1,N-p-1)\right\}, \tag{3.22}$$

if one neglects to condition on the fact that $Z=y_i$ was chosen to maximize one of the influence functions.

3.3 Model Checking

Bayesian analysis is certainly most effective when it is applied to deciding or inferring which single model of an exhaustive set of models is most appropriate for a given goal or the most appropriate mixture of the models. For example, if models M_1 and M_2 with accompanying parameter sets α_1 and α_2 and prior probabilities q_1 and q_2 $(q_1+q_2=1)$, are being entertained for observed data set $Y=y$, then the calculation of the ratio of

$$q_1 f(y|M_1) \;/\; q_2 f(y|M_2), \tag{3.22}$$

where

$$f(y|M_i) = \int f(y|M_i,\alpha_i)\; p(\alpha_i)\; d\alpha_i,$$

is appropriate for the comparison of M_1 and M_2. If prediction of a future observation is at issue then the calculation of the mixture is inferentially relevant

$$f(y_{n+1}|y) = q_1 f(y_{n+1}|y,M_1) + q_2 f(y_{n+1}|y,M_2). \tag{3.23}$$

Quite often either the models entertained are not exhaustive or one model conceptually appears, for certain reasons, pre-eminent in its explanatory or predictive potential and no other alternative model is entertained until this model's adequacy is sufficiently doubted.

Box (1980a, 1980b) working more or less along these lines devised an elegant approach to criticism of an entertained model. Assume that a single model M generating Y, given parameter set α, is structured such that

$$p(y,\alpha|M) = f(y|M,\alpha)\, p(\alpha|M) \tag{3.24}$$

where $f(y|M,\alpha)$is a joint probability function of Y conditional on the parameter set α, specified by M and $p(\alpha|M)$ is the prior probability function of α. The marginal probability function of Y given the model M is

$$p(y|M) = \int p(y,\alpha|M)\, d\alpha. \tag{3.25}$$

Then Box asserts, that by referring y to $p(y|M)$ as in

$$\gamma = \Pr\left\{Y: p(Y|M) < p(y|M)\right\} \tag{3.26}$$

or to some predictive checking function, say, g(Y) and referring g to $p(g|M)$

$$\gamma = \Pr[g(Y):\ p(g(Y)|M) < p(g(y)|M)] \tag{3.27}$$

that a small value of γ is indicative of a tentative inadequacy of the model. A number of useful applications of this marginal predictive significance test were also presented. There is no doubt that this is a highly useful procedure. However, one must realize that both marginal probability functions, $p(y|M)$ and $p(g|M)$ given by (3.24) and (3.25) depend in a certain sense on the sampling distribution of Y or g(Y). This particularly implies

that in cases of optional stopping or censoring where the likelihood remains unaltered, the sampling distribution of the observables can critically depend on the stopping rule or the censoring mechanism, neither of which are inherently a component of the model that requires checking. Hence this type of model criticism can be confounded with the stopping rule or the type of censoring, which may have little to do with criticism of the model. A particular case of this is examined for Bernoulli sampling by Geisser (1983). In this situation it is not clear as to what aspect of the model is called in to question other than the stopping rule.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. Bernoulli trials with probability of success θ, and uniform prior probability for θ. Then for a fixed number n of trials where y successes are observed, the predictive probability function of Y is easily calculated to be

$$\Pr(y|M) = \frac{1}{n+1} \qquad y = 0,1,\ldots,n. \tag{3.28}$$

i.e. uniform for all admissible values of y. Hence no test of the type

$$\Pr\{p(Y|M) < p(y|A)\} = \gamma \tag{3.29}$$

is available. Apparently predictive model criticism fails here.

If the experiment were terminated as soon as y successes were attained and resulted in n trials being observed, the predictive probability function of the number of trials is

$$\Pr(N = n|M) = \frac{y}{n(n+1)} \qquad n = y, y+1, \ldots \quad . \tag{3.30}$$

The fact that the probability function is monotonically decreasing in n indicates that the Box procedure is now available i.e. if the observed $N = n_0$ is large enough relative to y, the model may be called into question. In fact,

$$\Pr[N \geq n_0] = \sum_{n=n_0}^{\infty} \frac{y}{n(n+1)} = \frac{y}{n_0} = \gamma \qquad (3.31)$$

where $\gamma = \hat{\theta}$, the MLE of θ. This implies that predictive model criticism here succeeds only for small $\hat{\theta}$. Sampling until a fixed number of failures is attained results in criticism increasing with $\hat{\theta}$. In either case the only aspect of the model that can presumably be called into question, other than the uniform prior, is the stopping rule. But this is absurd.

The major difficulty here is the strong dependence of the marginal predictive distribution on the stopping rule itself given the model of i.i.d. Bernoulli trials with a uniform prior. So, in essence whatever is to be criticized is confounded with the stopping rule. Hence caution must be exercised when using the Box procedure in certain situations.

In this connection it is to be noted that predictive probability functions denoted in (3.20) and (3.21) are not susceptible to this criticism — they do not contradict the likelihood principal.

3.4 Illustration

The data in Table 1 graphed in Figure 1, from Aitchison and Dunsmore (1975, p. 182), is used to illustrate some of the techniques discussed in the previous parts of this section. Here a simple linear regression of y on x is fitted and the relevant PIF and estimative influence measures are tabled.

Table 1 Water contents (percentages by weight) of 16 soil specimens determined by two methods and associated influence measures.

Serial no. of specimen	Laboratory methods	On-site method	$2\hat{I}_i(Z)$	$2I(\beta,\sigma^2)$	$2\hat{I}_i(\beta)$	$\frac{v_i}{1-v_i}$	t_i^2
1	35.3	23.7	.22	.31	.24	.11	.18
2	27.6	20.2	.05	.00	.01	.07	.01
3	36.2	24.5	.21	.31	.25	.13	.17
4	21.6	15.8	.04	.01	.03	.08	.03
5	39.8	29.2	.08	.04	.06	.30	.01
6	24.1	17.8	.04	.00	.02	.07	.03
7	16.1	10.1	.07	.03	.04	.21	.00
8	27.5	19.0	.05	.00	.01	.07	.01
9	33.1	24.3	.06	.01	.02	.12	.00
10	12.8	10.6	1.09	1.18	.82	.19	.31
11	23.1	15.2	.04	.01	.03	.09	.03
12	19.6	11.4	.15	.26	.23	.17	.13
13	26.1	19.7	.03	.02	.04	.07	.05
14	19.3	12.7	.06	.01	.02	.13	.00
15	18.8	12.6	.07	.01	.02	.13	.00
16	39.8	31.8	.95	1.37	1.16	.48	.24

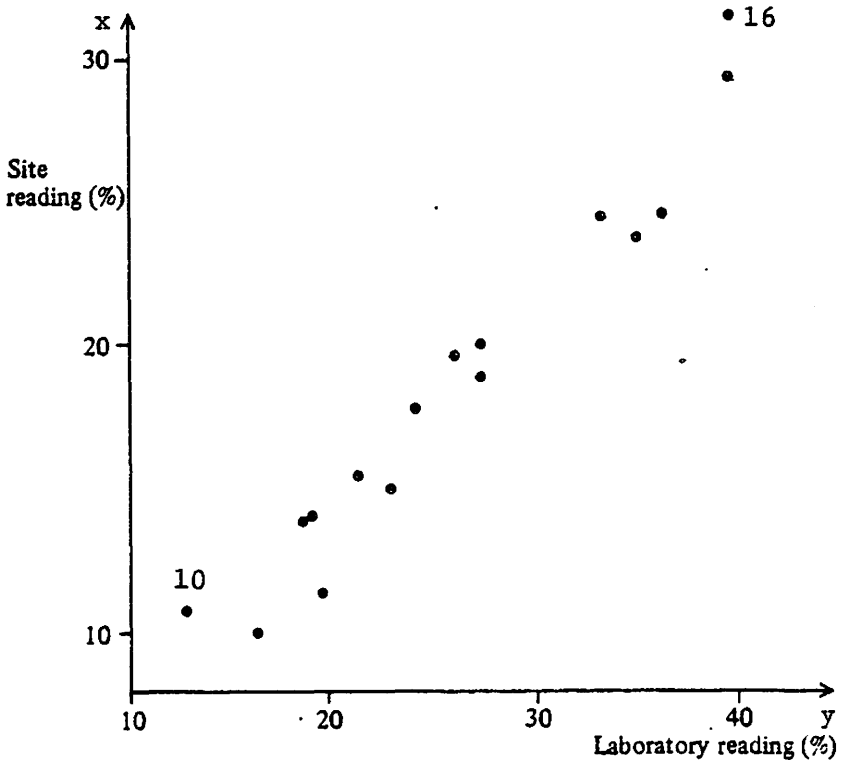

Fig. 1 Scatter diagram of laboratory and on-site measurements of 16 soil specimens.

Specimens 10 and 16 are denoted on the graph of Fig. 1 because they are the most influential on all three measures. Specimen 10 is most influential from the point of view of prediction while specimen 16 is most influential for the estimation of β alone or β and σ^2 jointly. However specimen 16 is influential mainly because it is the most distantly observed, while specimen 10 is far less so. But with respect to being consistent with the rest of the data, specimen 10 is most deviant but it is well within the range of acceptability. In fact the predictive probability that the maximum of (3.22) exceeds .31/.69 must be reasonably large, since ignoring that condition yields

$$\Pr[F(1,14) \geqslant .45] \doteq .51. \qquad (3.32)$$

For this simple regression situation, the graph itself reveals much the same information as the analysis of influential cases. It is clear that the model provides an adequate fit of the data and the most influential cases are those that are most distantly observed but are consonant with the model. Such techniques however are much more useful in multiple and multivariate regression situations, where subsets of data are to be examined for their influence and graphical displays are unavailable.

4. The Probability That A Future Fraction Will Lie In A Given Set.

A situation that often occurs in experimental work is where a sample of units is drawn from some "large" population of units and is measured with respect to some attribute or on a response to an administered agent. Before being measured the units are assumed indistinguishable with regard to the response (for the simplest paradigm) but they will inherently vary on the response, with the variation being natural and not an error of measurement (though measurement error may be an insignificant portion of the total variation). In this situation, inference and decision are often relevant for the response on a single future unit or several of them jointly or some special function of them depending on the goals of the investigation. For example, if an individual is about to take a new therapy whose response has been recorded for some experimental group of patients with whom this individual is presumed to be more or less indistinguishable with regard to the response, he would be interested only in the predictive distribution of a single future response (his own) or his chance of achieving some threshhold value. A physician treating M such patients would be more concerned with the fraction of them that exceed the critical threshhold. On the other hand, government health authorities who must deal with the possibility of an M which is a very large number and possibly not precisely known would be interested in the limiting fraction that exceeds the threshhold or varying threshholds. The latter case is one situation where the limiting value of the function (which is a parameter, in the sense that it is potentially unobservable) may be of interest. Other situations may arise where the distribution for moderate or large M is rather complex or intractable. In such situations the distribution of the limiting value may serve as a convenient approximation. Sometimes an informative summary of sorts is wanted for the response

but no particular fixed number of future values is of critical interest. In such a case one might want to focus on $M = 1$ and $M \to \infty$. Even in this situation it would be clearly more informative to present a whole spectrum of values for M. Traditionally, most statistical analyses focus on the infinite case and in so doing make statements only about parameter values. Further, even when making a statement about a parameter, the parameter should be the limit of a sensible function of future observables. The predictivistic point of view is that a statistical model is introduced not because it is necessarily the "true" one - it surely isn't - but because it will serve as an adequate approximation. What is most often critical then is not the fictive parameters of the convenient and approximate formulation but the potential observables.

Consider the set of random variables $X = (X^{(N)}; X_{(M)})$ where $X^{(N)} = (X_1, \ldots, X_N)$ are values that will be observed in the experiment and $X_{(M)} = (X_{N+1}, \ldots, X_{N+M})$ are values to be predicted. Assume that the joint probability function of X is

$$f(x^{(N)}; x_{(M)} | \alpha) = f(x_{(M)} | x^{(N)}, \alpha)\ f(x^{(N)} | \alpha) \qquad (4.1)$$

where α is a set of unknown parameters. For a given prior probability function $p(\alpha)$, we obtain the posterior density of α

$$p(\alpha | x^{(N)}) \propto f(x^{(N)} | \alpha)\ p(\alpha) \qquad (4.2)$$

and

$$f(x_{(M)} | x^{(N)}) = \int f(x_{(M)} | x^{(N)}, \alpha)\ p(\alpha | x^{(N)})\ d\alpha \qquad (4.3)$$

is the predictive probability function of the future set $X_{(M)}$ given $X^{(N)} = x^{(n)}$.

Many problems in statistics are formulated such that the sequence X_i, $i = 1,\ldots,N+M$ are independent and identically distributed. In such cases $X_{(M)}$ represents a set of exchangeable random variables so that each component, X_{N+i}, of $X_{(M)}$ has the same marginal distribution. Incidentally, because

$$f(x_{N+i}|x^{(N)}) = E_\alpha[f(X_{N+i}|\alpha)] \tag{4.4}$$

then from the point of view posterior squared error, $f(x_{N+i}|x^{(N)})$ is a "best" estimate of the common sampling distribution, Geisser (1971). Quite often one may be interested in a function (possibly vector valued) $g(X^{(M)})$ of the future values. Typically it may be the fraction of the observations that lie in some set I, say. Hence let

$$Y_i = \begin{cases} 1 & \text{if } X_{N+i} \in I \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

and $R = Y_1 + \ldots + Y_M$, then $\bar{Y} = RM^{-1}$ represents the fraction of future values that lie in I. Further, set $\theta(\alpha) = Pr(X_i \in I|\alpha)$, then via a simple conditioning argument

$$Pr[\bar{Y} = (r/M)] = \binom{M}{r} \int_0^1 \theta^r (1-\theta)^{M-r}\, p(\alpha|x^{(N)})\, d\alpha \tag{4.6}$$

When all the requirements of deFinetti's (1937) representation theorem are satisfied we also obtain

$$\lim_{M\to\infty} \bar{Y} = \theta \tag{4.7}$$

where θ is a random variable with posterior probability function $p(\theta|x^{(N)})$ derivable from $p(\alpha|x^{(N)})$.

The first two moments are also easily obtained,

$$E(\bar{Y}) = \int_I f(x_{N+i} | x^{(N)})\, dx_{N+i} = q \tag{4.8}$$

$$\mathrm{Var}(\bar{Y}) = q(1-q)\left[\frac{1-\rho}{M}+\rho\right] \tag{4.9}$$

where the common correlation coefficient is

$$\rho = [\Pr[Y_i = 1, Y_j = 1] - q^2]/q(1-q) \tag{4.10}$$

for $i \neq j$.

As an application, suppose we are dealing with a random sample from the translated exponential distribution where

$$F(x|\alpha,\gamma) = \begin{cases} 1 - e^{-\alpha(x-\gamma)} & \alpha > 0,\ x > \gamma > -\infty \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

Assume that $X_1,\ldots,X_d$ are fully observed values while $X_{d+1},\ldots,X_N$ are censored at $x_{d+1},\ldots,x_N$ respectively. Let

$$\begin{aligned} \bar{x}_d &= d^{-1}(x_1 + \ldots + x_d) \\ m_d &= \min(x_1,\ldots,x_d) \end{aligned} \tag{4.12}$$

and assume that

$$m_d \leq \min(x_{d+1},\ldots,x_N). \tag{4.13}$$

This latter condition that there is no censored value less than the minimum of the fully observed values, is very often met in practice and greatly simplifies the likelihood function and the presentation of subsequent formulas, see Geisser (1982a). There is no inherent difficulty, however, in allowing for the contrary of the above case except that the formulas become more complex because of their piecewise nature.

Further, we assume a conjugate prior density for α, γ to be

$$p(\gamma|\alpha) = N_0 \alpha e^{\alpha N_0 (\gamma - m_0)} \qquad \text{for } \gamma < m_0 \tag{4.14}$$

$$p(\alpha) = [N_0(\bar{x}_0 - m_0)]^{d_0-1} \alpha^{d_0-2} e^{-\alpha N_0(\bar{x}_0-m_0)}/\Gamma(d_0-1) \tag{4.15}$$

where $\alpha > 0$, $\bar{x}_0 > m_0$, and $1 < d_0 \leq N_0$ to insure that the distributions are proper.

From the above we can calculate

$$p(\gamma) = \frac{(d_0-1)(\bar{x}_0 - m_0)^{d_0-1}}{(\bar{x}_0 - \gamma)^{d_0}} \tag{4.16}$$

and

$$p(\alpha|\gamma) = [N_0(\bar{x}_0 - \gamma)]^{d_0} \alpha^{d_0-1} e^{-\alpha N_0(\bar{x}_0 - \gamma)}/\Gamma(d_0). \tag{4.17}$$

Suppose we define the survival function as

$$\theta = \Pr[Z > z|\alpha,\gamma] = \begin{cases} e^{-\alpha(z-\gamma)} & \alpha > 0,\ z > \gamma > -\infty \\ 1 & \text{otherwise.} \end{cases} \tag{4.18}$$

then one can obtain the posterior distribution of θ. For $0 < \theta < 1$

$$\Pr[\Theta \leq \theta | x^{(N)}] = \begin{cases} \theta^{N^*}\left(\dfrac{\bar{x}^* - m^*}{\bar{x}^* - z}\right)^{d^*-1} & \text{for } z \leq m^* \\ \theta^{N^*}\left(\dfrac{\bar{x}^* - m^*}{\bar{x}^* - z}\right)^{d^*-1} G\left(2N^*\left(\dfrac{\bar{x}^*-z}{m^*-z}\right)\log\theta\right) + 1 - G\left(2N^*\left(\dfrac{\bar{x}^*-m^*}{m^*-z}\right)\log\theta\right) & \text{for } m^* < z < \bar{x}^* \\ \theta^{N^*}\left(\dfrac{\bar{x}^* - m^*}{z - \bar{x}^*}\right)^{d^*-1} G\left(2N^*\left(\dfrac{z-\bar{x}^*}{m^*-z}\right)\log\theta\right) + 1 - G\left(2N^*\left(\dfrac{\bar{x}^*-m^*}{m^*-z}\right)\log\theta\right) & \text{for } z \geq \bar{x}^* \end{cases} \tag{4.19}$$

and

$$\Pr[\Theta = 1 | x^{(N)}] = \begin{cases} 1 - \left(\dfrac{\bar{x}^* - m}{\bar{x}^* - z}\right)^{d^*-1} & z < m^* \\ 0 & z \geq m^* \end{cases} \tag{4.20}$$

where G(u) represents the distribution function of a χ^2 variate with $2d^*-2$ degrees of freedom and

$$\begin{aligned} d^* &= d_0 + d \\ N^* &= N_0 + N \\ \bar{x}^* &= (N_0\bar{x}_0 + N\bar{x})/N^* \\ m^* &= \min(m_0, m). \end{aligned} \tag{4.21}$$

For the next observation X_{N+1} we can calculate the expectation of $\Pr[X_{N+1} > z|\alpha,\gamma]$ w.r.t. $p(\alpha,\gamma|x^{(N)})$ which results in the predictive survival function

$$\Pr[X_{N+1} > z] = \begin{cases} \dfrac{(N^*)^{d^*}(\bar{x}^* - m^*)^{d^*-1}}{(N^*+1)\,[z - m^* + N^*(\bar{x}^* - m^*)]^{d^*-1}} & z > m^* \\ 1 - (N^*+1)^{-1}\left(\dfrac{\bar{x}^*-m^*}{\bar{x}^*-z}\right)^{d^*-1} & z \leqslant m^* \end{cases} \tag{4.22}$$

Using (4.6), we obtain, Geisser (1982 b)

$$\Pr(\bar{Y} = \frac{r}{M}|z) = \begin{cases} \left(\dfrac{\bar{x}^* - m^*}{\bar{x}^* - z}\right)^{d-1} \binom{N^*+r-1}{r} \binom{N^*+M}{M} & r < M,\ z < m^* \\ 1 - \dfrac{M}{N^*+M}\left(\dfrac{\bar{x}^*-m^*}{\bar{x}^*-z}\right)^{d^*-1} & r = M,\ z < m^* \\ N^*\binom{M}{r}\sum_{j=0}^{M-r}\binom{M-r}{j}\dfrac{(-1)^j}{(N^*+r+j)}\left(1+\dfrac{(r+j)(z-m^*)}{N^*(\bar{x}^*-m^*)}\right)^{-(d^*-1)} & m^* < z \end{cases} \tag{4.23}$$

We further note, for the special limiting case of the conjugate prior on α and γ, namely the "noninformative" quasi prior which is

$$p(\alpha,\gamma) \propto \alpha^{-1}, \tag{4.24}$$

that the *'s are removed in (4.22) so that $d^* \to d$, $N^* \to N$, $\bar{x}^* \to \bar{x}$ and $m^* \to m$.

In cases where γ is known, by a simple translation, we effectively set $\gamma = 0$ and use (4.17) as the prior density for α. In this case

$$\Pr[\bar{Y} = \frac{r}{M} | z] = \binom{M}{r} (N^* \bar{x}^*)^{d^*} \sum_{j=0}^{M-r} \binom{M-r}{j} (-1)^j [N^* \bar{x}^* + z(r+j)]^{-d^*} \qquad (4.25)$$

and similarly for the vague prior in this instance

$$p(\alpha) \propto \alpha^{-1}, \qquad (4.26)$$

$d^* \to d$, $N^* \to N$ and $\bar{x}^* \to \bar{x}$, Geisser (1982a).

For large M, it is shown, Geisser (1982a) that the third expression of (4.23) and (4.25) can each be reasonably well approximated by replacing the Chi-squared distributions in (4.19) by F distributions with appropriate degrees of freedom.

When there is no censoring and lack of knowledge as to the common sampling distribution of $X_1, \ldots, X_N$, the type of calculation given in (4.6) could be the basis for a reasonably robust Bayesian procedure. For example, if we assume a uniform prior distribution for θ, in the Bayes-Laplace tradition, then we obtain

$$\Pr[\bar{Y} = \frac{r}{M} | z] \doteq \frac{\binom{r+s}{s}\binom{M+N-r-s}{N-s}}{\binom{M+N+1}{M}}, \qquad (4.27)$$

where s is the number of X_i's, $i = 1, \ldots, N$ that exceed z. It is to be recalled that the first use of (4.6) was the calculation of (4.27) made by Laplace (1774) where $p(\theta | x^{(N)})$ was based on a uniform prior for θ as alluded to in the introduction.

This is a much coarser method since it sacrifices the finer distinctions engendered by shifts in z and the distributional attributes of θ.

Calculating (4.6) explicitly for many distributions is often difficult. For example, consider the simple normal case i.e. $X_1,\ldots,X_{N+M}$ are $N(\mu,1)$ without censoring. In this case

$$\theta = Pr[Z > z\,|\mu] = 1 - \Phi(z-\mu) \tag{4.28}$$

where $\Phi(\cdot)$ is the standard normal distribution function. In the simple situation where the quasi-prior for μ is uniform so that the posterior distribution for μ is $N(\bar{x},N^{-1})$, the calculation of (4.6)

$$\sqrt{\frac{N}{2\pi}}\binom{M}{r}\int[1-\Phi(z-\mu)]^r\,\Phi^{M-r}(z-\mu)e^{-\frac{N}{2}(\mu-\bar{x})^2}\,d\mu \tag{4.29}$$

requires a series expansion and consequently an approximation. An alternative way of handling this problem is to calculate the joint predictive density of the set $X_{(M)}$,

$$f(x_{(M)}\,|x^{(N)}) = \int \varphi(N^{\frac{1}{2}}(\bar{x}-\mu))\ \prod_{i=1}^{M}\varphi(x_{N+i}-\mu)\ d\mu \tag{4.30}$$

where $\varphi(\cdot)$ is the standard normal density. Hence the predictive distribution of the components of $X_{(M)}$ is exchangeable being a multivariate normal distribution with common mean $\bar{x}$, common variance $1+M^{-1}$ and common covariance M^{-1}. If we now require the probability that exactly r out M lie in the same interval i.e. $X_{N+i} > z$, this still remains a formidable calculation. The complexity increases when we permit σ^2 to be unknown and use the simple quasi-prior

$$p(\mu,\sigma^2) \propto 1/\sigma^2.$$

Now we obtain for the joint distribution of $X_{N+1},\ldots,X_{N+M}$ an exchangeable

multivariate student distribution. This multivariate student distribution can be reasonably well-approximated by an exchangeable multivariate normal distribution with the same mean but slightly inflated variances and covariances. This has the advantage that an approximate solution for the more complex case depends on an exact solution for the simpler case.

In either case the probability that exactly r out of the next M observations will exceed z can be represented by

$$P_r = \binom{M}{r} \Pr[X_{N+1} > z, \ldots, X_{N+r} > z, \; X_{N+r+1} \leq z, \ldots, X_{N+M} \leq z]$$

because the variables are exchangeable multivariate normal or student variables. Exact calculations for P_r seem prohibitive, but approximation or, perhaps, bounds similar to the type developed by Kounias (1968) and Hunter (1976) may be useful here.

4.1 Examples

In this section we present two examples illustrating the methods obtained for the exponential distribution.

Example 1.

A department store's past experience with a type of flourescent light is given in days to failure as follows:

7	26	35	49	69	99	141
15	27	36	56	71	105	145
20	28	41	57	75	106	154
21	29	43	62	78	126	168
22	34	48	64	91	133	189

with 5 lights exceeding 196 days.

A new room in the store is being opened up and will require 10 lights. The manager wants to have an idea of the lifetime distribution of these 10 lights. Assuming an exponential survival distribution with the lower limit known to be 0 and using the quasi-prior of (4.26), we calculate

$$\Pr\left[\bar{Y} \leqslant \frac{r}{10} \,\middle|\, z\right]$$

for $r = 0,1,\ldots,10$; $z = 7, 14, 21, 28,\ldots,196$. This should provide adequate information concerning the lifetimes of the 10 lights (see table 2).

Table 2

$\Pr[\bar{Y} \leq \frac{r}{10} | z]$ for flourescent light data by weeks $w = z/7$

r \ w	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.01	.01	.02	.03	.04	.05	.06	.07	.09	.11	.13	.15	.17	.20	.21	.24
1	.00	.00	.00	.00	.00	.00	.00	.00	.01	.02	.03	.04	.06	.08	.10	.13	.16	.20	.23	.27	.31	.35	.39	.43	.47	.50	.54	.57
2	.00	.00	.00	.00	.00	.01	.01	.03	.05	.07	.10	.14	.18	.23	.28	.33	.38	.44	.49	.54	.58	.63	.67	.70	.74	.77	.80	.82
3	.00	.00	.00	.00	.01	.03	.06	.09	.14	.20	.26	.32	.39	.45	.52	.58	.63	.68	.73	.77	.80	.83	.86	.88	.90	.92	.93	.94
4	.00	.00	.01	.02	.05	.10	.16	.23	.31	.40	.48	.55	.62	.68	.74	.78	.82	.86	.88	.91	.93	.94	.95	.96	.97	.98	.98	.98
5	.00	.01	.03	.08	.16	.25	.35	.45	.54	.63	.70	.76	.81	.85	.89	.91	.93	.95	.96	.97	.98	.98	.99	.99	.99	.99	1.00	1.00
6	.00	.04	.11	.22	.35	.47	.58	.68	.76	.82	.87	.90	.93	.95	.96	.97	.98	.99	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	.03	.14	.30	.46	.60	.72	.80	.86	.91	.94	.96	.97	.98	.99	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	.15	.39	.59	.74	.84	.90	.94	.96	.98	.99	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	.50	.75	.87	.93	.97	.98	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mode $(\bar{Y})$	1.00	.90	.80	.80	.70	.70	.60	.60	.50	.50	.50	.40	.40	.40	.30	.30	.30	.30	.20	.20	.20	.20	.20	.10	.10	.10	.10	.10
Median $(\bar{Y})$	.90	.90	.80	.80	.70	.70	.60	.60	.50	.50	.50	.40	.40	.40	.30	.30	.30	.30	.30	.20	.20	.20	.20	.20	.20	.10	.10	.10
Mean $(\bar{Y})$	.93	.87	.81	.75	.70	.65	.61	.57	.53	.50	.46	.43	.40	.38	.35	.33	.31	.28	.27	.25	.23	.22	.20	.19	.18	.17	.15	.14

Example 2: The data set in the following table is reported in Pike (1966) and discussed by Kalbfleisch and Prentice (1980). The table gives the times from insult with a carcinogen to death for two differentially treated groups.

Table 3 Days to Vaginal Cancer Mortality in Rats

Group 1	143,	164,	188,	188,	190,	192,	206,	209,	213,	216,
	220,	227,	230,	234,	246,	265,	304,	216*,	244*	
Group 2	142,	156,	163,	198,	205,	232,	232,	233,	233,	233,
	233,	239,	240,	261,	280,	280,	296,	296,	323,	204*,
	344*									

*Censored.

Assuming a known lower level of 100 days, Kalbfleisch and Prentice use the exponential distribution to compare the groups and determine that there is no difference in survival between the two groups but indicated that the exponential fit was an inadequate description of the data. Pike (1966) found that the third power of the excess over 100 days was adequate to provide an exponential fit. Hence we assume $U = (X - 100)^3$ is exponentially distributed with known minimum for U to be 0 and compute the predictive survival probability $Pr[X_{N+1} > x]$ using the quasi-prior of (4.24) for the two groups in Table 4 which is plotted in Figure 2. In Table 5 we present a brief comparison of the probabilities that the fraction of all future rats will survive beyond a varying threshhold for the two groups. All of the computations indicate the survival superiority of group 2, assuming the adequacy of the model.

Table 4

Predictive probability that a random rat will survive x days when treated either as in group 1, $X_{1,N+1}$ or as in group 2, $X_{2,N+1}$.

x	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
$Pr[X_{1,N+1} > x]$	.95	.91	.86	.80	.72	.64	.55	.47	.38	.30	.23	.17	.12	.09	.06	.04
$Pr[X_{2,N+1} > x]$	.97	.95	.92	.89	.84	.78	.73	.67	.60	.53	.45	.39	.32	.26	.21	.16

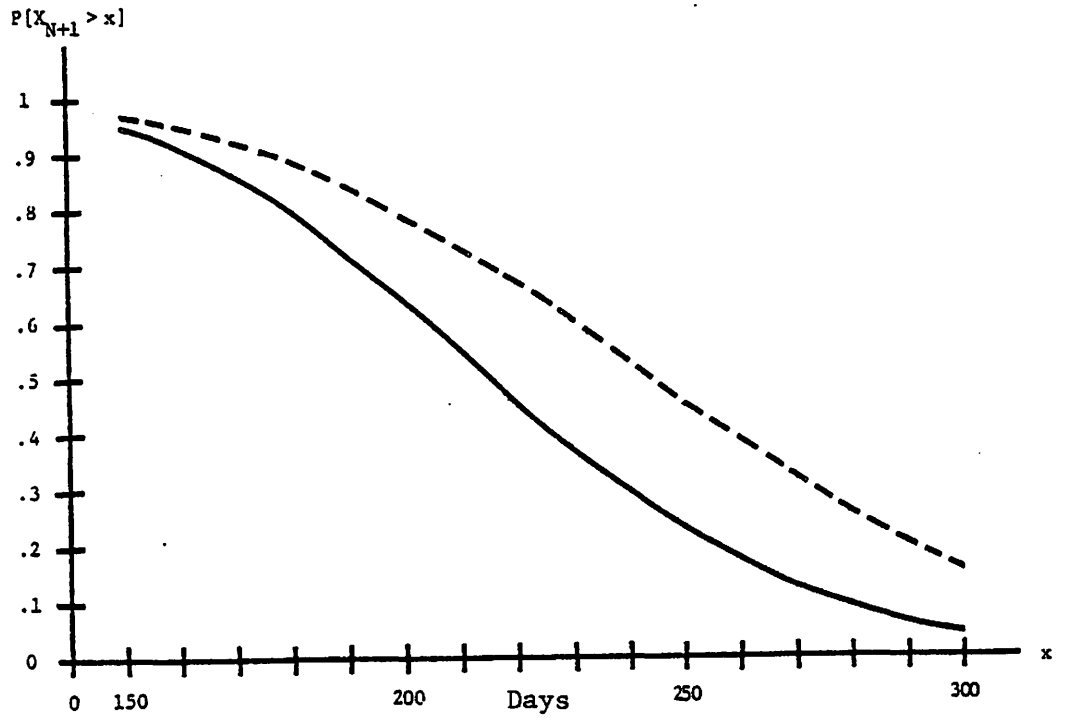

Fig. 2

Comparison of survival probabilities for two groups of rats with differential carcinogenic results.

Table 5

$Pr[\theta > \theta | x]^*$: Probability that at least a given fraction of all future rats survive beyond a specified threshhold x for groups 1 and 2.

θ \ x	200		225		250		275		300	
	1	2	1	2	1	2	1	2	1	2
.1	1.00	1.00	1.00	1.00	.97	1.00	.45	1.00	.05	.82
.2	1.00	1.00	1.00	1.00	.62	1.00	.07	.87	.00	.26
.3	1.00	1.00	.93	1.00	.20	.97	.01	.43	.00	.04
.4	1.00	1.00	.60	1.00	.03	.74	.00	.10	.00	.00
.5	.98	1.00	.19	.97	.00	.04	.00	.01	.00	.00
.6	.73	1.00	.02	.69	.00	.00	.00	.00	.00	.00
.7	.20	.97	.00	.15	.00	.00	.00	.00	.00	.00
.8	.01	.42	.00	.00	.00	.00	.00	.00	.00	.00
.9	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
E(θ)	.64	.78	.42	.63	.23	.45	.10	.29	.04	.16

*An entry of 1.00 in this table indicates that the value is between 1 and .995. An entry of .00 indicates that the value is between 0 and .005.

5. Acknowledgment.

I am indebted to Murray Clayton and Wesley Johnson for computational assistance.

References

Aitchison, J. and I. R. Dunsmore (1975). *STATISTICAL PREDICTION ANALYSIS*, Cambridge University Press, New York, Melbourne .

Aitchison, J. and Sculthorpe, Diane (1965). Some problems of statistical prediction. *BIOMETRIKA* 52, 469-83.

Akaike, H. (1978). A new look at the Bayes procedure, *BIOMETRIKA*, 65, 1, 53-59.

Anderson, T. W. (1971). *THE STATISTICAL ANALYSIS OF TIME SERIES*, John Wiley and Sons. New York.

Andrews, D., and Pregibon D. (1978). Finding outliers that matter, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, Ser. B, 40, 85-93.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *REGRESSION DIAGNOSTICS*, New York: John Wiley.

Box, G.E.P. (1980a). Sampling and Bayes' inference in scientific modelling and robustness. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, A, 143, 383-430.

Box, G.E.P. (1980b). Sampling inference, Bayes' inference, and robustness in the advancement of learning. *BAYESIAN STATISTICS*, ed. J.M. Bernardo et al. 366-370.

Box, G.E.P., and M. Jenkins (1970). *TIME SERIES ANALYSIS FORECASTING AND CONTROL*, Holden-Day, Inc., San Francisco.

Butler, R. and E.D. Rothman (1975). Intervals based on reuse of the sample, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 75, 372, 881-889.

Condorcet, N.C. (1786). Réfexious sur la méthod de déterminer la probabilité des évènemens futurs, apres l'observation des evenemens passés. *MÉMOIRS DE L'ACADÉMIE ROYALE DES SCIENCES*, 539-559.

Cook, R.D. (1977). Detection of influential observations in linear regression, *TECHNOMETRICS*, 19, 15-18.

————(1979), Influential observations in linear regression, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 74, 169-174.

Cook, R.D., and Weisberg, S. (1980), Finding influential cases in regression: a review, *TECHNOMETRICS*, 22, 495-508.

————(1982). *RESIDUALS AND INFLUENCE IN REGRESSION*, Chapman and Hall, New York.

deWaal, D. et al (1981) Model selection, prediction and estimation for multivariate normal populations, *JOURNAL OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION*, 15, pp. 113-127.

References

Dunsmore, I.R. (1968). A Bayesian approach to calibration. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, B 30, 396-405.

————(1969). Regulation and optimization. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, B,31, 160-70.

————(1974). The Bayesian predictive distribution in life testing models. *TECHNOMETRICS*, 16, 455-60.

Fearn, T. (1975). A Bayesian approach to growth curves. *BIOMETRIKA*, 62 1 89-100.

Fisher, R.A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters, *PROCEEDINGS OF THE ROYAL SOCIETY*, A, 139, 343-348.

————(1934). Probability likelihood and quantity of information in the logic of uncertain inference, *PROCEEDINGS OF THE ROYAL SOCIETY*, A, 146, 1-8.

————(1956). *STATISTICAL METHODS AND SCIENTIFIC INFERENCE*, Oliver & Boyd, Edinburgh .

Geisser, S. (1964). Posterior odds for multivariate normal classification. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, Ser. B 26, 69-76.

————(1965). Bayesian estimation in multivariate analysis. *ANNALS OF MATHEMATICAL STATISTICS*, 36, 150-159.

————(1966). Predictive discrimination. In: P. Krishnaiah, ed., *MULTIVARIATE ANALYSIS*, Academic Press, New York, 149-163.

————(1970). Bayesian analysis of growth curves. *SANKHYĀ*, Ser. A 32, 53-64.

————(1971). The inferential use of predictive distributions. V. Godambe and D. Sprott, eds., *FOUNDATIONS OF STATISTICAL INFERENCE*, Rinehart and Winston, 456-469.

————(1974). A predictive approach to the random effect model. *BIOMETRIKA*, 61, 101-107.

————(1975a). The predictive sample reuse method with applications. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 70, 320-328, 350.

————(1975b). A new approach to the fundamental problem of applied statistics. *SANKHYĀ*, 37, B, 385-397.

References

————(1976). Review of statistical prediction analysis, by J. Aitchison and I.R. Dunsmore, *BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY*, Sept., 683-688.

————(1979). Discussion on: Reference posterior distribution for Bayesian inference, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, B, 136-137.

————(1980a). Sample reuse selection and allocation criteria, *MULTIVARIATE ANALYSIS V*, edited by P. Krishnaiah, North Holland Publishing Co., 387-398.

————(1980b). Predictive sample reuse techniques for censored data, *BAYESIAN STATISTICS* (with discussion) University Press, Valencia, Spain, 430-468.

————(1980c). The estimation of distribution functions and the prediction of future observations, *PROCEEDINGS OF THE CONFERENCE ON RECENT DEVELOPMENTS IN STATISTICAL METHODS AND APPLICATIONS*, Academia Sinica, Taiwan, 193-208.

————(1981). Sample reuse procedures for predicting the unobserved portion of a partially observed vector, *BIOMETRIKA*, 243-250.

————(1982a). Aspects of the predictive and estimative approaches in the determination of probabilities. *BIOMETRICS*, 38, Supplement, March, 75-93 (with discussion).

————(1982b). Predicting Pareto and exponential observables. *UNIVERSITY OF MINNESOTA* Technical Report 408.

Geisser, S. and Desu, M.M. (1968). Predictive zero-mean uniform discrimination. *BIOMETRIKA*, 55, 519-24.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection, *JOURNAL OF AMERICAN STATISTICAL ASSOCIATION*,]979, Vol. 74, No. 365, 153-160. Corregendum, 1980, Vol. 75, No. 371, 765.

Good, I.J. (1965). *THE ESTIMATION OF PROBABILITIES: AN ESSAY ON MODERN BAYESIAN METHODS*, MIT Press.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, B 29, 83-100.

————(1970). *STATISTICAL TOLERANCE REGIONS: CLASSICAL AND BAYESIAN*. London: Griffin.

Guttman, I. and Tiao, G.C. (1964). A Bayesian approach to some best population problems. *ANNALS OF MATHEMATICAL STATISTICS*, 35, 825-35.

References

Hoaglin, D.C., and Welsch, R.E. (1978), The hat matrix in regression and ANOVA, *THE AMERICAN STATISTICIAN*, 32, 17-22, and "Corrigenda," 145.

Hunter, D. (1976). An upper bound for the probability of a union, *JOURNAL OF APPLIED PROBABILITY*, 13, 597-603.

Jaynes, E.T. (1980). The minimum entropy production principle. *ANNUAL REVIEW OF PHYSICAL CHEMISTRY*, 31, 579-601.

Jeffreys, H. (1932). On the theory of errors and least squares, *PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON*, Ser. A, 138, 48-55.

————(1933). Probability, statistics, and the theory of errors, *PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON*, Ser. A, 140, 523-535.

————(1934). Probability and scientific method, *PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON*, Ser. A, 146, 9-15.

————(1939). *THEORY OF PROBABILITY*, Oxford: Claredon Press.

Johnson, W., and Geisser, S. (1982). Assessing the predictive influence of observations, *STATISTICS AND PROBABILITY ESSAYS IN HONOR OF C.R. RAO*, edited by Kallianpur, Krishnaiah, and Ghosh, Amsterdam, North Holland, 343-358.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 78, 381, 137-144.

————(1981). Estimative influence functions for the multivariate general linear model. Tech. Report No. 23, University of California, Davis.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *THE STATISTICAL ANALYSIS OF FAILURE TIME DATA*, John Wiley and Sons, New York.

Kounias, E. (1968). Bounds for the probability of a union with applications. *ANNALS OF MATHEMATICAL STATISTICS*, 39, 2154-2158.

Kullback, S., and Leibler, R.A. (1951). On information and sufficiency, *ANNALS OF MATHEMATICAL STATISTICS*, 22, 79-86.

Laplace, P.S. (1774). Memoir sur la probabilité des causes par les évènemen. *MÉMOIRS DE L'ACADÉMIE ROYALE DES SCIENCES*, 6, 621-656.

Pearson, K. (1907). On the influence of past experience on future expectation, *PHILOSOPHICAL MAGAZINE*, XII ser. 6, 365-378.

————(1920). The fundamental problem of practical statistics, *BIOMETRIKA*, 13, 1-16.

————(1979). *THE HISTORY OF STATISTICS IN THE 17th AND 18th CENTURIES*, ed. Pearson, E.S., London, Griffin.

References

Pike, M.C. (1966). A suggested method of analysis of a certain class of experiments in carcinogenesis, *BIOMETRICS*, 22, 142-161.

Roberts, H.V. (1965). Probabilistic prediction, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 60, 309, 50-62.

Stigler, S.M. (1982). Thomas Bayes's Bayesian inference. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, A, 145, 2, 250-258.

Stone, M. (1974a). Cross-validatory choice and assessment of statistical predictions (with discussion) *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, B36, 111-147.

————(1974b). Cross-validation and multinomial prediction, *BIOMETRIKA*, 61, 509-515.

————(1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY*, Ser. B, 39, 44-47.

Wahba, G. (1977). Optimal smoothing of density estimates. *CLASSIFICATION AND CLUSTERING*, Academic Press, New York. 423-458.

Whittle, P. (1963). *PREDICTION AND REGULATION BY LINEAR LEAST SQUARES METHODS*, VanNostrared, Princeton, N.J.

Wiener, N. (1949). *THE EXTRAPOLATION, INTERPOLATION AND SMOOTHING OF STATIONARY TIME SERIES WITH ENGINEERING APPLICATIONS*, Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass.

Wilks, S.S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *ANNALS OF MATHEMATICAL STATISTICS*, 13, 400-9.

Wold, H. (1938). *A STUDY IN THE ANALYSIS OF STATIONARY TIME SERIES*, Almquist and Wicksell. Stockholm.

————(1959). Ends and means in econometric model building, *PROBABILITY AND STATISTICS: THE HAROLD CRAMÉR VOLUME*, ed. Grenander, U. John Wiley and Sons, New York. 355-434

Yaglom, A.M. (1962). *AN INTRODUCTION TO THE THEORY OF STATIONARY RANDOM FUNCTIONS*, (Revised English Edition, Richard A. Silverman, trans. and ed.), Prentice-Hall, Inc., Englewood Cliffs, N.J.